

优达学城数据分析师纳米学位

A/B 测试项目

说明: [点击此处下载此文档的英文版本](#)。

模板格式

此模板可用于组织你最终项目的答案。应从你的答案中复制到小测试中的项用蓝色显示。

试验设计

指标选择

列出你将在项目中使用的不变指标和评估指标。(这些应与你在“选择不变指标”和“选择评估指标”小测试中使用的指标一样)

不变指标: 选用 Number of cookies 和 Number of clicks

评估指标: 选用 Gross Conversion 和 Net Conversion

对于每个指标,解释你为什么使用或不使用它作为不变指标或评估指标。此外,说明你期望从评估指标中获得什么样的试验结果。

选择 Number of cookies 及 Number of clicks 作为不变指标,主要是因为这两个指标不会因为试验的进行而发生变化。其余指标均有可能因为 Experiment 组中新增的页面而发生变化。

选择 Gross conversion 和 Net conversion 作为评估指标,因为本试验的目的是评估新增的页面是否会使得那些没有足够时间学习课程的用户停止注册,进而能够集中资源为有足够时间学习的用户提供更好的体验和辅导。在所有指标中能够最直接评估该项影响的是 Gross conversion、Retention 和 Net conversion 三项指标。当分组单元为事件时,根据分析方法得到的差异性可能很接近根据经验计算的差异性,所以分组单元选择 Number of clicks 而不是 Number of cookies。Gross conversion 和 Net conversion 的分析单元均为 Number of clicks,而 Retention 的分析单元为 Number of user-ids,若选用此指标易造成分析方法与经验分析差异性过大,若时间允许,可采用经验方法计算该指标。我在后续试验中舍弃了 Retention,保留 Gross conversion 和 Net conversion 作为评估指标。

期望从评估指标中获得的试验结果:通过设立对照试验观察分析两组的 Gross conversion 和 Net conversion 是否出现显著差异。若试验组的 Gross conversion 显著低于控制组,则表明新增的页面能够有效阻止学习时间不足的用户进行注册;若试验组的 Net conversion 显著高于控制组,则表明新增的页面能够使得有时间完成课程和项目且学习意愿强烈的用户留下。

测量标准偏差

列出你的每个评估指标的标准偏差。(这些应是来自“计算标准偏差”小测试中的答案。)

Gross conversion 的 Standard deviation 为 0.0202。

Net conversion 的 Standard deviation 为 0.0156。

对于每个评估指标，说明你是否认为分析估计与经验变异是类似还是不同（如果不同，在时间允许的情况下将有必要进行经验估计）。简要说明每个情况的理由。

本实验选择的分组单元是 Number of clicks, 评估指标 Gross conversion 及 Net conversion 的分析单元均为 Number of clicks, 所以分析估计与经验计算的差异性匹配的。

规模

样本数量和功效

说明你是否会在分析阶段使用 Bonferroni 校正，并给出实验正确设计所需的页面浏览量。（这些应是来自“计算页面浏览量”小测试中的答案。）

不会在分析阶段使用 Bonferroni 校正，因为 Gross conversion 和 Net conversion 具有关联性，使用 Bonferroni 校正得到的结果过于保守。

使用 $\alpha=0.05$, $(1-\beta)=0.8$ 。

Gross conversion: $d_{min}=0.01$, Baseline conversion rate = 20.625%, 根据在线计算器得到一组实验所需 Gross conversion 的样本量为 25835, 两组实验所需的样本量为 51676, 然后转化为需要的页面浏览量 = $51676/0.08 = 645950$

Net conversion: $d_{min} = 0.075$, Baseline conversion rate = 10.93125%, 根据在线计算器得到一组实验所需 Net conversion 的样本量为 27413, 两组实验所需的样本量为 54826, 然后转化为需要的页面浏览量 = $54826/0.08 = 685325$

考虑到所需页面浏览量要同时覆盖两个指标，因此选择两个页面浏览量中较大的一个，即 685325

持续时间和曝光比例

说明你会将多少百分比的页面流量转入此试验，以及鉴于此条件，你需要多少天来运行试验。（这些应是来自“选择持续时间和曝光”小测试中的答案。）

一般来说，A/B testing 的实验时间为一个月左右，根据所需的页面浏览量（685325）及每天的页面浏览量（40000），曝光的流量部分为 50%时所需天数为 35 天，符合要求。因此选择曝光比例为 50%，试验持续时间为 35 天。

说明你选择所转移流量部分的原因。你认为此试验对优达学城来说有多大风险？

不选择对所有流量开展实验的原因主要是出于以下几个方面：

一是安全性。推出这个新的页面弹窗，我们不确定它是否能在所有浏览器中正常运行，也不确定用户将有什么反应，所以选择仅向部分用户开展实验。

二是来自外部压力。我们不能确定这个新的页面弹窗是否会被保留，当我们不能确定这个事项时也不希望人们看到它后在社交网站上大肆宣扬。

三是随机分配分组单元时，为了避免异常数据（例如节假日等等）对试验结果的误导，也倾向于压缩发送流量比例，在合理范围内延长持续时间，从而尽可能的了解用户在不同日

期甚至是一天中不同时间段的差异。

四是有可能同时进行多个实验。

对此实验，优达学城面临的风险相对较大，一是新的弹窗页面不一定能在所有浏览器中正常运行；二是用户有可能在社交网站上宣传优达学城对课程学习时间的要求，从而使得一部分有意向的潜在用户望而却步等。

试验分析

合理性检查

对于每个不变指标，对你在 95%置信区间下期望观察到的值、实际观察的值及指标是否通过合理性检查给出结论。（这些应是来自“合理性检查”小测试中的答案）

Number of cookies: 控制组总计有 345543 个观测值，试验组总计有 344660 个观测值。对于每一个观测值，它被发送至控制组和试验组的几率均为 50%，且事件之间相互独立，因此符合二项分布特征。在 $\alpha = 0.05$ 水平下， $SE = ((0.5*0.5) / (N-con + N-exp))^{0.5} = 0.0006$ ， $margin\ of\ error = SE * Z-score = 0.0006 * 1.96 = 0.0012$ ，围绕 0.5 为中心的置信区间为(0.4988,0.5012)。实际观测值= $N-con / (N-con + N-exp) = 0.5006$ ，位于置信区间之内，因此 Number of cookies 通过 Sanity check。

Number of clicks: 控制组总计有 28378 个观测值，试验组总计有 28325 个观测值。对于每一个 click 事件，它被发送至控制组和试验组的几率均为 50%，且事件之间相互独立，因此符合二项分布特征。在 $\alpha = 0.05$ 水平下， $SE = ((0.5*0.5) / (N-con + N-exp))^{0.5} = 0.0021$ ， $margin\ of\ error = SE * Z-score = 0.0021 * 1.96 = 0.0041$ ，围绕 0.5 为中心的置信区间为(0.4959,0.5041)。实际观测值= $N-con / (N-con + N-exp) = 0.5005$ ，位于置信区间之内，因此 Number of clicks 通过 Sanity check。

所有不变指标均通过 Sanity check。

结果分析

效应大小检验

对于每个评估指标，对试验和对照组之间的差异给出 95% 置信区间。说明每个指标是否具有统计和实际显著性。（这些应是来自“效应大小检验”小测试的答案。）

Gross conversion:

$$P-pool = (X-con + X-exp) / (N-con + N-exp) = (3785+3423) / (17293+17260) = 0.2086$$

$$SE-pool = (P-pool*(1-P-pool)*(1/N-con + 1/N-exp))^{0.5} = 0.0044$$

$$Margin\ of\ error = SE-pool * Z-score = 0.0086$$

$$d-hat = P-exp - P-con = X-exp/N-exp - X-con/N-con = -0.0206$$

$$CI: (-0.0291, -0.0120)$$

置信区间不包括 0，具有统计显著性，位于最小实质显著性左边界-0.01 左方，具有实质显著性。

Net conversion:

$$P-pool = (X-con + X-exp) / (N-con + N-exp) = (2033+1945) / (17293+17260) = 0.1151$$

$$SE-pool = (P-pool*(1-P-pool)*(1/N-con + 1/N-exp))^{0.5} = 0.0034$$

$$Margin\ of\ error = SE-pool * Z-score = 0.0067$$

$d\text{-hat} = P\text{-exp} - P\text{-con} = X\text{-exp}/N\text{-exp} - X\text{-con}/N\text{-con} = -0.0049$

CI: (-0.0116, 0.0019)

置信区间包括 0，不具有统计显著性，位于最小实质显著性右边界 0.0075 左方，不具有实质显著性

符号检验

对于每个评估指标，使用每日数据进行符号检验，然后报告符号检验的 p 值以及结果是否具有统计显著性。（这些应是“符号检验”小测试中的答案。）

Gross conversion:

将每日试验组与控制组的 Gross conversion 做差，为正值的有 4 天，全部观测天数为 23 天，使用在线计算器得到双尾 p 值为 0.0026，小于 $\alpha = 0.05$ ，因此具有统计显著性。

Net conversion:

将每日试验组与控制组的 Net conversion 做差，为正值的有 10 天，全部观测天数为 23 天，使用在线计算器得到双尾 p 值为 0.6776，远大于 $\alpha = 0.05$ ，因此不具有统计显著性。

汇总

说明你是否使用了 Bonferroni 校正，并解释原因。若效应大小假设检验和符号检验之间存在任何差异，描述差异并说明你认为导致差异的原因是什么。

未使用 Bonferroni 校正。因为 Gross conversion 与 Net conversion 具有关联性，而 Bonferroni 校正在此种情况下过于保守。

效应大小假设检验与符号检验之间未存在差异。

建议

提供建议并简要说明你的理由。

建议进行后续试验。因为评估指标 Gross conversion 的效应大小假设检验和符号检验均显示，此项更改极有可能通过阻止那些时间未达要求的用户学习课程降低了用户的注册数量。但 Net conversion 的效应大小检验和符号检验却未能说明此项更改是否确实提高了注册用户的付费率。因此建议设计后续试验进一步检验如何提升用户的付费情况。

后续试验

对你会开展的后续试验进行概括说明，你的假设会是什么，你将测量哪些指标，你的转移单位将是什么，以及做出这些选择的理由。

在此试验的基础上，对于那些参加了 14 天免费课程却点击取消的用户增加弹窗，询问是否愿意接受更为专业的教练辅导和项目审阅服务，并在弹窗中附往期毕业学员成功就职名

企的简述和详细页面链接。若点击是，提示学员将在 14 天免费试用到期后自动扣款，若点击否，则关闭弹窗确认取消课程。

假设：参加免费课程的学员有可能因为课程内容难度过大而放弃，也有可能因为学习课程后能否顺利就业等存有疑虑。通过提示用户付费后能够接受更为专业的辅导和服务及往期毕业学员成功就职名企的信息，能够有效打消拟取消课程用户的疑虑，增加试用用户付费比例。

测量指标包括：

1. 参加 14 天免费课程的 user-ids 数量
2. 14 天免费课程结束后自动扣费的 user-ids 数量
3. 付费率：即 14 天免费课程结束后自动扣费的 user-ids 数量占参加 14 天免费课程的 user-ids 数量的比例

转移单位：由于以上测量指标全部为 user-ids 的数量，所以转移单位自然选择 user-ids 的数量。